MULTI-BEAM AND MULTI-TASK LEARNING FOR JOINT SOUND EVENT DETECTION AND LOCALIZATION

Technical Report

Wei Xue, Ying Tong*, Chao Zhang, Guohong Ding*

JD AI Research

ABSTRACT

Joint sound event detection (SED) and sound source localization (SSL) is essential since it provides both the temporal and spatial information of the events that appear in an acoustic scene. Although the problem can be tackled by designing a system based on the deep neural networks (DNNs) and fundamental spectral and spatial features, in this paper, we largely leverage the conventional microphone array signal processing techniques to generate more comprehensive representations for both SED and SSL, and to perform post-processing such that stable SED and SSL results can be obtained. Specifically, the features extracted from signals of multiple beams are utilized, which orient towards different directions of arrival (DOAs), and are formed according to the estimated steering vector of each DOA. Smoothed cross-power spectra (CPS) are computed based on the signal presence probability (SPP), and are used both as the input features of the DNNs, and for estimating the steering vectors of different DOAs. A triple-task learning scheme is developed, which jointly exploits the classification and regression based criterion for DOA estimation, and uses the classification based criterion as a regularization for the DNN. Experimental results demonstrate that the proposed method yields substantial improvements compared with the baseline method for the task 3 of the DCASE challenge 2019.

Index Terms— Sound event detection, Sound source localization, Deep neural networks

1. INTRODUCTION

Sound event detection (SED) aims to determine the time period that an acoustic event appears, and been widely used in applications such as robotics, smart home and surveillance [1–3]. Recently the focus has shifted to not only estimating the temporal information of the acoustic event, but also determining the location of the corresponding sound source. This raises the problem of joint SED and sound source localization (SSL), namely, sound event localization and detection (SELD). A microphone array is usually utilized, such that the temporal and spatial samplings are simultaneously performed to provide a richer description of the acoustic scene.

Conventionally the SED and SSL are separately treated, and there have been enormous researches on each problem. Most of the state of the art SED systems are now based on deep neural networks (DNN) [1, 3–8], and the convolutional neural networks (CNN) [6, 7, 9] and recurrent neural networks (RNN) [5, 6] are exploited to learn the compact representations and the temporal characteristics of the acoustic events, respectively. For SSL, conventional methods are generally based on analysing the cross-correlations between the multichannel signals [10–15], including the generalized cross-correlation (GCC) [10], the multichannel cross-correlation coefficient (MCCC) [16] and the multiple signal classification (MUSIC) [17–19] based approaches. Methods based on the DNNs are also proposed [20–24], which use the cross-correlations as the input feature, and estimate the direction of the arrival (DOA) as a regression or classification problem.

Since both SED and SSL can be achieved by using a DNN, to cope with the SELD problem, a DNN-based end-to-end system is usually trained, and the SED and SSL results are simultaneously obtained by multi-task learning [25, 26]. Shared bottom hidden layers are used to extract features for both SED and SSL, and then the layers are split into different branches to adapt to the specific task. In [26], the baseline system of the DCASE 2019 SELD task, the conventional recurrent neural network (CRNN) is used to extract the spectral and temporal characteristics of the acoustic events, and the SED and SSL branches are respectively formulated using the feed-forward (FF) networks.

In this paper, the conventional microphone array signal processing is largely leveraged to generate comprehensive representations for both SED and SSL, and to performing post-processing. We extract the features from beamformed signals for multiple DOAs, and compute the smoothed cross-power spectra (CPS) based on the signal presence probability (SPP). The SELD problem is solved in a triple-task learning scheme that uses the classification based SSL criterion as a regularization for the DNN. Experimental results demonstrate the superior performance of the proposed method compared with the baseline system.

The rest of the paper is organized as follows. In Section 2 we present the signal model. Feature extraction and the DNN structure will be introduced in Section 3 and 4 respectively. The methods for postprocessing the DNN outputs are described in Section 5, and Section 6 elaborates the data augmentation and system ensemble strategies. We evaluate the proposed method in Section 7 and draw conclusions in Section 8.

2. SIGNAL MODEL

We briefly introduce the signal model that will facilitate the derivations in this paper. With $Q <= 2$ sources and $M = 4$ microphones, the short-time Fourier transform (STFT) domain reverberant signal in the m-th microphone is expressed as

$$Y_m(t,f) = \sum_{q=1}^{Q} H_{m,p}(f) S_q(t,f), \tag{1}$$

where $S_q(t,f)$ is the STFT signal of the q-th source, $H_{m,q}(f)$ is the room impulse response (RIR) from the q-th source to the m-th

*Equal contribution.

microphone. We ignore the additive noise since the signal-to-noise ratio (SNR) is typically high in the training and testing sets of the DCASE 2019 SELD task.

3. FEATURE EXTRACTION

The steps of extracting the features by using the conventional signal processing techniques are introduced in this section.

3.1. Smoothed CPS

The phase of the CPS implies the phase difference between channels, which is closely related to the DOA of the sound source. With the multichannel signal, the CPS between the m-th microphone and the first microphone is computed by recursive smoothing as

$$R_{m1}(t,f) = \alpha R_{m1}(t-1,f) + (1-\alpha)Y_m^*(t,f)Y_1(t,f), \quad (2)$$

where $0 < \alpha < 1$ is a smoothing factor.

Since the phase difference between channels is randomly distributed when the signal is inactive, we apply a mask to the CPS, such that the effect of inactive periods is removed, and the mask is computed according to the signal presence probability (SPP) using the first channel signal based on [27]. With the SPP, $P(t,f)$, the time-frequency (TF) domain mask is defined by

$$\mathcal{M}(t,f) = \{1, \text{ if } P(t,f) > 0.6; 0, \text{ otherwise.}\} \quad (3)$$

and the resulting masked CPS is calculated as

$$\tilde{R}_{m1}(t,f) = R_{m1}(t,f)\mathcal{M}(t,f). \quad (4)$$

Since a hard thresholding is performed in (3), to further remove the fluctuations caused by the probably non-smooth mask, the smoothed CPS is further obtained by

$$\bar{R}_{m1}(t,f) = [\tilde{R}_{m1}(t-1,f) + \tilde{R}_{m1}(t,f)]/2. \quad (5)$$

3.2. Steering Vector

The steering vector describes the relationships between the multichannel signals for a certain DOA. Although in this challenge the explicit positions of the microphones are unknown, it is possible to estimate the steering vectors for different DOAs from the labelled development dataset.

Ignoring the level difference, the $M \times 1$ steering vector is determined by the phase differences between channels, whose m-th element can be calculated from the extracted smoothed SPS by

$$G_m(f,\phi) = \mathbb{E}\{\frac{\bar{R}_{m1}(t,f)}{|\bar{R}_{m1}(t,f)|}\}, \text{where the DOA in } (t,f) \text{ is } \phi. \quad (6)$$

With the development dataset, for each DOA, the segments with only one active source are selected, and the corresponding steering vector is computed by averaging the phase of the smoothed CPS in this segments. Since 36 azimuths and 9 elevation angles are present, totally $36 \times 9 = 324$ steering vectors are computed.

Fig. 1 displays the angular spectrum of the steering vector between the first two channels. The DOA index is divided into 36 intervals and in each interval 9 elevations are included. It can be seen that the angular spectrum changes smoothly for different elevations in each interval, and the angular spectrum can act as a distinct feature for each DOA. The estimated steering vector will be used to perform beamforming and to estimate the DOA of each source in the subsequent procedures.

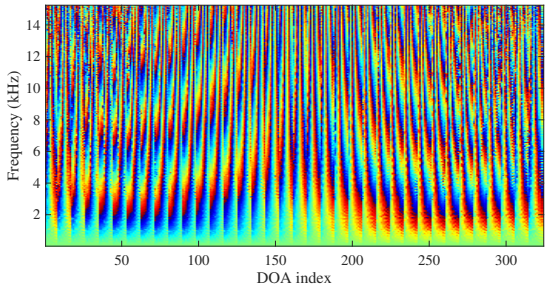

Figure 1: The angular spectrum of the steering vector between the second and first channel.

3.3. Multiple DOA Beamforming

With the estimated steering vectors, we use beamforming to achieve signal separation in the case of overlapping acoustic events. Since the DOA of each source is unknown, eight fixed beamformers are designed to extract the signals from different DOAs, and the two elevations and four azimuths of the eight target DOAs are $\{-25^\circ, 25^\circ\}$ and $\{-170^\circ{:}90^\circ{:}100^\circ\}$, respectively. Simple delay and sum (DS) beamformers are adopted, since the noise level is low, and can be assumed to be spatially white across microphones (In this case the minimum variance distortion response (MVDR) beamformer is equivalent to the DS beamformer). In this way eight beamformed signals are obtained, and used to extract the power spectrum features along with the multichannel raw signals.

3.4. Power Spectrum Features

We extract the power spectrum features of the twelve channel signals, which include the four channel raw observations and eight channel beamfomer outputs. 96-dimensional log-Mel and constant Q-transform (CQT) features are extracted based on the STFT-domain power spectra, and are fed into the DNN for SELD.

4. DNN FOR SELD

4.1. Input Feature

The input features of the DNN consist of three parts: 1) the log-Mel and CQT features described in Section 3.4, 2) the angular spectra of the four-channel STFT-domain microphone signals, 3) the angular spectra of the four-channel smoothed CPS. The last two parts represent the phase information that is useful for SSL. Here the FFT size is chosen as 2048, and only the lower 512 bins of the whole frequency range are used to compute the angular spectra of the STFT-domain signals and the smoothed CPS.

4.2. DNN structure

The structure of the DNN is shown in Fig. 2. In the proposed network structure, separate CNNs are adopted for the log-Mel features and the phase (angular spectrum) features respectively, and the learned features are concatenated and sent into the GRU layers to account for the temporal evolution of the acoustic event. Different with the baseline method in [26] which jointly learns a classification network for SED and a regression network for SSL, here, a triple-task learning scheme is developed, which jointly exploits the clas-

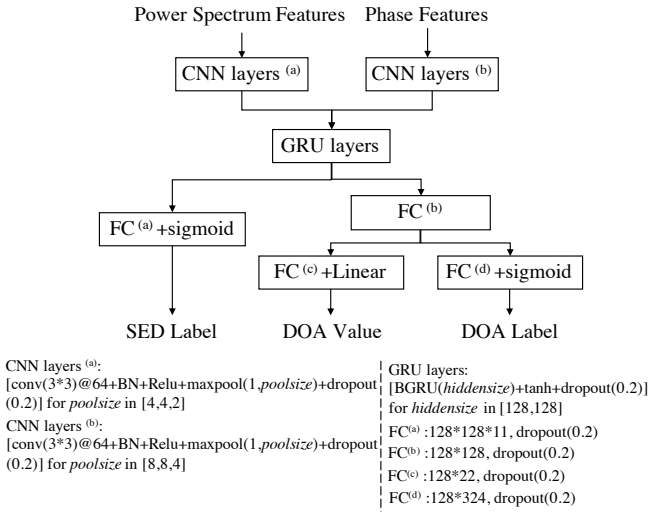

Figure 2: The DNN structure for SELD. BN represents batch normalization, and FC represents fully-connected network.

sification and regression based criterion for DOA estimation. In the classification based DOA estimation network, since $36*9$ DOAs are considered in the 3D space, the learning target is defined as a 324 sparse vector, whose element is one if the sound source appears at the corresponding DOA. The classification based DOA estimation criterion acts as a regularization for the DNN, and it is shown that it helps to improve the performances of both SED and SSL, which are determined by the SED label and DOA value. The optimization criterion for the SED classification, DOA classification and DOA regression are binary cross-entropy, binary cross-entropy and mean square error, respectively, and the loss functions are combined with a weight as [1,50,50] for joint optimization during training.

It should be noted that the DOA cannot be inferred only from the DOA label outputs because of the ambiguity of assigning the DOA to the correct sound source in the overlapping scenarios. Therefore, the regression based DOA learning target is necessary for constructing the DNN.

5. POST-PROCESSING

5.1. DOA Estimation

Although the SSL is integrated into the DNN as a learning task, the resulting DOA estimates exhibit high fluctuation over time. This is probably because that, although the phase features for SSL keep constant during the source active period, since the DNN is also adapted for SED, the layers for SSL are effected by the time-varying log-Mel features. To overcome this problem, the DOA is finally estimated by the conventional SRP-PHAT method, and the DOA estimates from the DNN are only used as anchor points to solve the ambiguity problem of assigning DOA to the source.

Given the steering vector and the smoothed CPS estimated in Section 3, in each frame, a 36×9 spatial spectrum can be computed, whose element indicates the cost function value of the SRP-PHAT algorithm. According to the number of active sources $\hat{Q}$ inferred by the SED estimation, the corresponding highest $\hat{Q}$ peaks are selected from the SRP-PHAT spatial spectrum, and the peak is determined if the point satisfies: a) being the highest point in a 5×5 neighbouring region which picks the neighbours by $-2 : 1 : 2$, and b) being higher than 0.7 times the highest SRP-PHAT value. If two source are active simultaneously, a peak is assigned to the closer source according to the DOA estimated by the DNN.

Table 1: Systems for ensemble

System	Signals for feature extraction
Log-Mel	4-channel raw signals
Log-Mel&CQT	4-channel raw signals
Log-Mel + BF	4-channel raw signals + 8-channel beamformer outputs

5.2. Temporal Spatial Consistency

We assume that a sound source is static and there is no short pause during the active period. Short pauses with duration less than 200 ms are removed by using the estimation results before the pause. Then the estimated DOAs for each continuous time interval of the acoustic event are further smoothed using a weighted voting strategy. For each frame in the active interval, a vote is given to the corresponding DOA, and two additional votes are assigned if there is only one active source in the current frame. The DOA of the acoustic event for the whole active interval is determined by the DOA with the highest votes, and then applied to all frames in the active interval.

6. DATA AUGMENTATION AND SYSTEM ENSEMBLE

The development data is augmented to improve the generalization ability of the trained model. Similar to automatic speech recognition [28], the speed of the training data is perturbed by 0.9 and 1.1, respectively, and the label of the speed-perturbed training data can be calculated by scaling the beginning and finishing time of each acoustic event by the corresponding factor. During training, we exploit the dataset with the original speed as the validation set to select the best model.

In the development stage, we develop three subsystems to finally produce an ensemble system. The three subsystems exploit the same DNN structure and post-processing strategy, and differentiates only in the power spectrum part of the input features, which are summarized in Table 1. In each frame, the SED estimate of the ensemble system is first determined by weighted majority voting, then the DOA is given by the DOA output of the system that yields the estimated event type. In the case of controversial DOA estimates from different systems, the system with smaller DOA difference between the DNN output and post-processing is chosen. Different ensemble weights can be chosen for each subsystem. During ensembling, the subsystems are duplicated according to their weights, and the SELD results are finally obtained from all the duplicated subsystems by majority voting. For instance, if the weight vector is [0.2,0.2,0.6], the third system is duplicated three times in the majority voting stage.

In the evaluation, since four cross-validation combinations are available for each system in Table 1, in total we exploit 12 systems to generate the SELD estimates for the evaluation set. The same strategy is utilized to ensemble the 12 systems.

Table 2: Performances on the development set

System	ER	F-score	DOA error	FR	SELD score
baseline	0.39	0.775	34.82	0.829	0.245
Log-Mel	0.215	0.869	9.36	0.888	0.127
Log-Mel&CQT	0.173	0.899	9.66	0.890	0.110
Log-Mel + BF	0.148	0.914	9.78	0.907	0.095
Xue_JDAI_task_3_1	**0.113**	**0.934**	**9.00**	**0.906**	**0.081**

Table 3: Ensemble weights for submitted systems

System	Weighting vector for Log-Mel, Log-Mel&CQT, Log-Mel+BF
Xue_JDAI_task_3_1	[0.33,0.33,0.33]
Xue_JDAI_task_3_2	[0.16, 0.33, 0.5]
Xue_JDAI_task_3_3	[0.2, 0.2, 0.6]
Xue_JDAI_task_3_4	[0.2,0.4,0.4]

7. EVALUATION

7.1. Data and Experimental Setup

The proposed method is evaluated on the DCASE 2019 challenge SELD development dataset [29, 30], which is recorded using a four channel microphone array. The development set is divided into four cross-validation splits, with each split consisting of 100 utterances sampled at 48 kHz, and the duration of each utterance is one minute. Eleven kinds of acoustic events are included, and up to two events may appear simultaneously in a frame. The azimuths and the elevations of the acoustic events are distributed within the range of $[-180^\circ, 170^\circ]$ and $[-40^\circ, 40^\circ]$, respectively, all with a 10° increment.

We compare the systems used for ensemble and the ensemble system with the baseline method in [26]. In all experiments, the hop size is set to 20 ms and the FFT size is 2048. Each minibatch contains 16 feature sequences, with each of them having a length of 128.

The metrics introduced on [31] are used for evaluation, which include the F-score and error rate (ER) for SED, and the DOA error and frame recall (FR) for SSL. An SELD score is calculated based on the above four metrics to provide an overall evaluation of the SELD performance.

7.2. Results

The performances of different systems are shown in Table 2. Only the ensemble system "Xue_JDAI_task_3_1" is shown for example. We totally generate four ensemble systems which adopt different weights for ensembling the subsystems. The weights for different ensemble systems are summarized in Table 3. It can be seen that all the three proposed systems yield substantially better results than the baseline system. Compared with the system which only adopts the Log-Mel features of the four-channel signals, we notice that, the absolute ER of SED is reduced by 4% by exploiting the CQT features, and another 2.5% gain is obtained by using the beamforming outputs from multiple DOAs. By integrating the three systems, we can finally achieve an ER of 11.3% for SED, which is significantly lower than the baseline system. On the other hand, it is shown that the DOA error is reduced to approximately 9° with the DOA estimation method in Section 5.1.

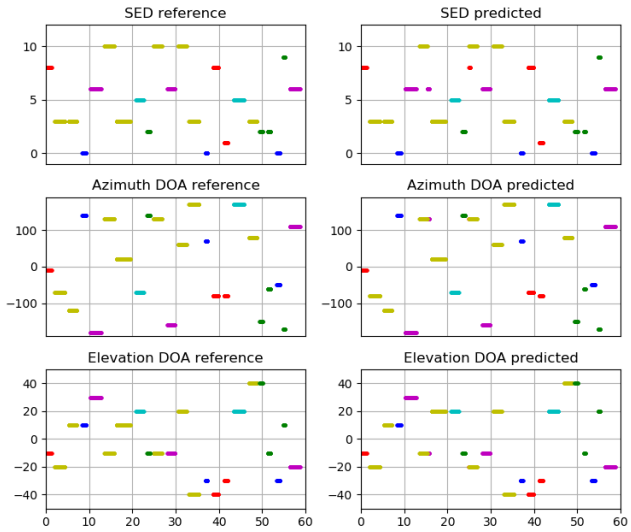

Figure 3: Illustration of the SELD performance for the ensemble system.

Fig. 3 displays the SELD performance of the ensemble system on one utterance. We can observe that for almost all cases the proposed system can yield both accurate and stable SED and SSL estimates.

8. CONCLUSIONS

In this paper we propose we an SELD method which leverage the conventional microphone array signal processing techniques. Based on the smoothed CPS, the steering vector for each DOA is estimated, and is used to design the beamformers for multiple DOAs and for DOA estimation in the postprocessing stage. A triple-task learning scheme is used, which uses the classification based SSL criterion as a regularization for the DNN. The effectiveness of the proposed method is demonstrated by the experiments on the development dataset of DCASE 2019 challenge SELD task.

9. REFERENCES

[1] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *Proc. Intl. Conf. Robotics and Automation*, 2018, pp. 74–79.

[2] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," *ACM Computing Surveys (CSUR)*, vol. 48, no. 4, p. 52, 2016.

[3] C. Grobler, C. P. Kruger, B. J. Silva, and G. P. Hancke, "Sound based localization and identification in industrial environments," in *IECON 2017-43rd Annual Conference of the IEEE Industrial Electronics Society*, 2017, pp. 6119–6124.

[4] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *Proc. Intl. Joint Conf. on Neural Networks (IJCNN)*, 2015, pp. 1–7.

[5] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real

life recordings," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6440–6444.

[6] S. Adavanne, P. Pertilä, and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 771–775.

[7] Y. Xu, Q. Kong, Q. Huang, W. Wang, and M. D. Plumbley, "Convolutional gated recurrent neural network incorporating spatial features for audio tagging," in *Proc. Intl. Joint Conf. on Neural Networks (IJCNN)*, 2017, pp. 3461–3466.

[8] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2015, pp. 1–6.

[9] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 559–563.

[10] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[11] Y. Huang and J. Benesty, "Adaptive multichannel time delay estimation based on blind system identification for acoustic source localization," in *Adaptive Signal Processing*. Springer Berlin Heidelberg, 2003, pp. 227–247.

[12] J. Tugnait, "Time delay estimation with unknown spatially correlated Gaussian noise," *IEEE Trans. Signal Process.*, vol. 41, no. 2, pp. 549–558, 1993.

[13] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP J. on Applied Signal Processing*, vol. Special issue on advances in multimicrophone speech processing, pp. 1–19, 2006.

[14] W. Xue, S. Liang, and W. Liu, "Interference robust DOA estimation of human speech by exploiting historical information and temporal correlation." in *Proc. Conf. of Intl. Speech Commun. Assoc. (INTERSPEECH)*, 2013, pp. 2895–2899.

[15] W. Xue, W. Liu, and S. Liang, "Noise robust direction of arrival estimation for speech source with weighted bispectrum spatial correlation matrix," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 5, pp. 837–851, 2015.

[16] J. Benesty, J. Chen, and Y. Huang, "Time-delay estimation via linear interpolation and cross correlation," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 509–519, Sept. 2004.

[17] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[18] J. Dmochowski, J. Benesty, and S. Affes, "Broadband MUSIC: Opportunities and challenges for multiple source localization," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2007, pp. 18–21.

[19] Y. Zhang and B. P. Ng, "MUSIC-like DOA estimation without estimating the number of sources," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1668–1676, Mar. 2010.

[20] K. Youssef, S. Argentieri, and J.-L. Zarader, "A learning-based approach to robust binaural sound localization," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pp. 2927–2932.

[21] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 2814–2818.

[22] N. Ma, G. J. Brown, and T. May, "Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions," in *Proc. Conf. of Intl. Speech Commun. Assoc. (INTERSPEECH)*, 2015.

[23] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 405–409.

[24] D. Salvati, C. Drioli, and G. L. Foresti, "Exploiting cnns for improving acoustic source localization in noisy and reverberant conditions," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 103–116, 2018.

[25] T. Hirvonen, "Classification of spatial audio location and content using convolutional neural networks," in *Audio Engineering Society Convention 138*. Audio Engineering Society, 2015.

[26] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE J. Sel. Topics Signal Process.*, pp. 1–1, 2018.

[27] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[28] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Conf. of Intl. Speech Commun. Assoc. (INTERSPEECH)*, 2015.

[29] S. Adavanne, A. Politis, T. V. Adavanne, A. Politis, and T. Virtanen, "TAU spatial sound events 2019 - ambisonic and microphone array, development datasets," *Zenodo*, 2019.

[30] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," in *Submitted to Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE)*, 2019.

[31] "DCASE 2019 sound event detection and localization," http://dcase.community/challenge2019/task-sound-event-localization-and-detection, accessed: 2019-06-10.